FDR doesn't tell the whole story: Joint influence of effect size and covariance structure on the distribution of the false discovery proportions.

Alan H. Feiveson[1], Ph.D., James Feidler[1,2], Ph.D., Robert J. Ploutz-Snyder[1,2], Ph.D.,PStat®

[1]NASA Johnson Space Center, Houston, TX [2]Universities Space Research Association, Houston, TX

Abstract

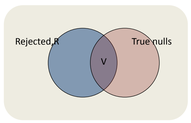
As part of a 2009 Annals of Statistics paper, Gavrilov, Benjamini, and Sarkar report results of simulations that estimated the false discovery rate (FDR) for equally correlated test statistics using a well-known multiple-test procedure. In our study we estimate the distribution of the false discovery proportion (FDP) for the same procedure under a variety of correlation structures among multiple dependent variables in a MANOVA context. Specifically, we study the mean (the FDR), skewness, kurtosis, and percentiles of the FDP distribution in the case of multiple comparisons that give rise to correlated non-central *t*-statistics when results at several time periods are being compared to baseline. Even if the FDR achieves its nominal value, other aspects of the distribution of the FDP depend on the interaction between signed effect sizes and correlations among variables, proportion of true nulls, and number of dependent variables. We show examples where the mean FDP (the FDR) is 10% as designed, yet there is a surprising probability of having 30% or more false discoveries. Thus, in a real experiment, the proportion of false discoveries could be quite different from the stipulated FDR.

Background and Significance

Gavrilov, Benjamini, and Sarkar (GBS) [1] discuss the pros and cons of several methods for controlling the FDR in a multiple-testing situation with a large number of variables which may be correlated. In particular, they prove that a simplified version of a family of multistage procedures suggested by Benjamini, Krieger, and Yekutieli (BKY) [2] does indeed control the FDR to a desired level *q*, when the test statistics are mutually independent. GBS then provide results of some simulations with equi-correlated and normally distributed test statistics to show that the FDR of this simplified BKY procedure is fairly robust under this dependence model.

False Discovery Proportion and Rate

R = # of hypotheses rejected
V = # of true null hypotheses rejected



False discovery proportion (FDP)

$$FDP = \begin{cases} V/R & (R > 0) \\ 0 & (R = 0) \end{cases}$$

FDR = E(*FDP*)

Study Summary

- 20, 40, 80, 160, 320 variables
- 1000 simulated experiments per simulation run
- dependence scenarios (DS):
  - I - variables and tests are completely independent;
  - PI - variables are independent, but with dependent multiple comparisons as a result of repeated measures design;
  - D - general covariance structure between variables (weighted sum of AR1, constant correlation, and two-stage Wishart) arising from multiple comparisons.
- FDP attributes studied: mean (FDR), median, IQR, skewness, kurtosis

Simplified BKY Procedure

As defined in [1] , the adaptive step-down procedure based on *m* tests is as follows:

1. Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the ordered *p*-values.
2. Define critical values α_i as follows:
   $$\alpha_i = iq/(m + 1 - i(1 - q))$$
3. Let *k* be defined by
   $$k = \max\{1 \le i \le m: p_{(j)} \le \alpha_j, j = 1, \ldots, i\}$$
4. If *k* exists, reject *k* hypotheses with p-values $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$; Otherwise reject no hypotheses.

Simulated Data Model

Longitudinal observations of n_v variables from *n* "subjects" at n_t times:

$$Y_{ij} \sim N_{n_v}(\mu_j, \rho R_i + (1 - \rho)R_{ij}) \quad (0 \le \rho \le 1)$$

For each variable *k*, test $H_{0k}: \mu_{kj} = \mu_{k1}$ with post-ANOVA contrasts

Total of $m = n_v(n_t - 1)$ *t*-statistics, each with $(n - 1)(n_t - 1)$ d.f.

Effect sizes: $\varepsilon_{kj} = \frac{\sqrt{n}|\mu_{kj} - \mu_{k1}|}{(1-\rho)\sqrt{2}}$

Covariance structure of $n_v n_t$ observations per "subject":

$$V = V_1 \times V_0$$
$$V_0 = (1 - \rho)I_{n_t} + \rho J_{n_t} \equiv R_{n_t}(\rho)$$
$$V_t = V(Y_{ij}) = w_1 AR1(\theta) + w_2 R_{n_t}(c) + w_3 S$$

Simulation Process

$n = 20$, $n_t = 4$, $n_v = 20, 40, 80, 160, 320$; FDR controlled to $q = 0.10$

$\rho, \theta, c \sim [U(0,1)]^{1/2}$ or $U(0,1)$; w's random weights

S = 2-stage sample correlation matrix

Matrix of mean vectors

$$\mu = [\mu_1|\mu_2|\mu_3|\mu_4] \text{ where } \mu_j \text{ is } n_v \times 1$$

Each row of μ has one of the following forms (at random)

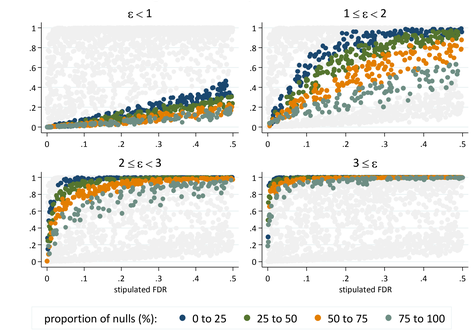
$$\mu = \begin{matrix} 1: \\ 2: \\ 3: \\ 4: \end{matrix} \begin{bmatrix} 0 & \Delta & \Delta & \pm\Delta \\ 0 & 0 & \Delta & \pm\Delta \\ 0 & 0 & 0 & \pm\Delta \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $p_z = m_0/m$, the expected proportion of nulls in the last three columns, is distributed as $U(0, 0.90)$ and the sign of Δ in the last column is $\pm$ 1 with probability 0.5.

The effect size $\varepsilon \sim U(0,4)$ and $\Delta = \frac{\varepsilon \cdot (1 - \rho)\sqrt{2}}{\sqrt{n}}$

Results

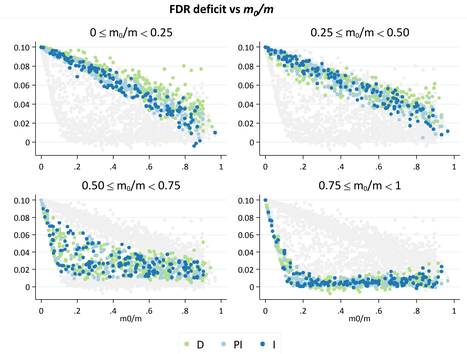
Average Power vs Stipulated FDR level



As expected, power increases with stipulated FDR and the increase is sharper for larger effect size and a greater percentage of non-nulls.

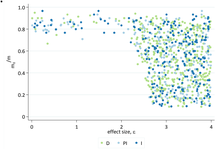
When does the FDR attain its nominal level?

FDR deficit vs m_0/m



The FDR deficit is defined as *q* – *FDR* where *q* is the desired FDR control setting (in this case *q* = 0.1 for all simulations). Note how the FDR is generally conservative in that it is often considerably smaller than *q*, except for when the proportion of nulls is large and the effect size is fairly large (say > 2).

In particular the combinations of m_0/m and ε such that the FDR is "close" to *q* (say |q – FDR| < .025) is quite striking (next Figure): Either m_0/m is > 0.8, or ε > 2.5. This effect appears to hold regardless of the dependence scenario.



Results (continued)

FDP Distributions and Effect Size, m_0/m, and Dependence

When the effect size and m_0/m are small, the FDP distribution can have extreme positive skewness, with a high probability of no rejections. In this case the FDR is also below its nominal value (larger FDR deficit). Skewness appears to increase under dependence.

When the effect size is large and m_0/m is small, skewness of the FDP distribution is relatively close to zero, but is largest in Dependence Scenario D. Imposition of multiple comparisons in the independent variables case (DS = PI) tends to produce negatively skewed FDP distributions. Under complete independence (DS = I), FDP skewness is small and can be positive or negative.

When m_0/m is large (> 0.5), FDP skewness is largest when DS = D (small or large effect size), but there is not much between Dependence Scenarios for moderate effect sizes. There is little difference in skewness between DS = I and DS = PI, over the range of effect sizes studied.

While there is little effect of DS on the FDR (previous figure), regardless of m_0/m or effect size. However increased skewness increases the probability of realizing a large FDP in any given experiment.

Unlike the case with skewness, the dependence scenario has little effect on the IQR of the FDP distribution, regardless of ε or m_0/m.

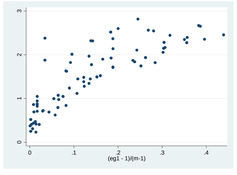
Continuous Effect of Dependence

A good continuous index of dependence is the normalized largest eigenvalue of the m x m correlation matrix of the t-statistics.

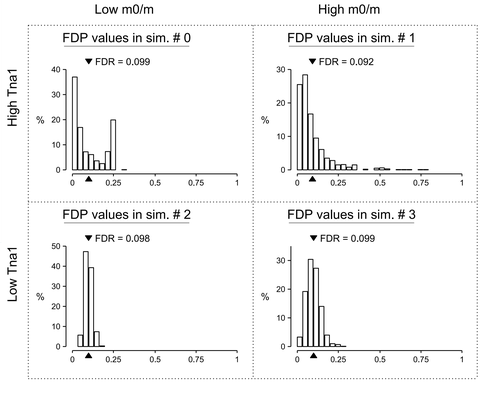
$$\lambda_1^* = \frac{\lambda_1 - 1}{m - 1}$$

For certain combinations of ε and m_0/m, many properties of the FDP distribution are strongly associated with λ_1^*.

Example: Skewness vs. λ_1^* for large ε and m_0/m



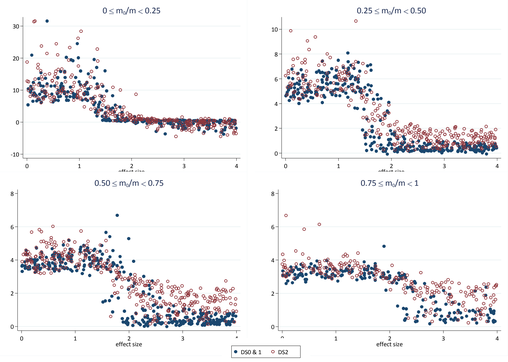
Example of FDP Distributions by Dependence



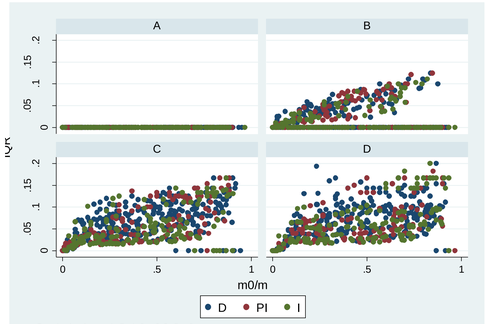
The best FDP distributions tend to arise when λ_1^* is small (little or no dependence) and especially when there is a small proportion of nulls. Here, the distribution is more symmetric about the nominal FDR (0.1), and also has relatively little spread (small IQR).

By contrast when test statistics are highly dependent (high λ_1^*), the FDP distributions can be bimodal or highly skewed with large spread.

Skewness of FDP Distribution



Inter-quartile Range (IQR)



Graphs by A: es<1; B: 1-2 c:2-3, D:>3

Association of Dependence with FDP Characteristics

Effect Size Range	Proportion of Nulls (%)	N	Association with Dependence Index (Somers' D): Mean (FDR)	Median	IQR	Skewness	Kurtosis
0-1	0-25	116	-0.06	0.00	0.00	0.09	0.11
	25-50	110	-0.31	0.00	0.00	0.30	0.30
	50-75	142	-0.40	0.00	0.00	0.37	0.38
	75-100	93	-0.43	0.00	0.00	0.41	0.40
1-2	0-25	111	0.12	-0.12	0.06	0.14	0.14
	25-50	124	-0.02	-0.23	-0.16	0.21	0.17
	50-75	105	-0.27	-0.20	-0.23	0.12	-0.03
	75-100	74	-0.36	-0.03	-0.07	0.24	0.11
2-3	0-25	103	-0.12	-0.22	0.38	0.26	0.17
	25-50	124	0.03	-0.35	0.52	0.71	0.57
	50-75	139	-0.16	-0.48	0.21	0.66	0.53
	75-100	75	-0.23	-0.44	-0.20	0.55	0.41
3-4	0-25	126	-0.14	-0.25	0.16	-0.09	-0.16
	25-50	124	0.19	-0.58	0.58	0.67	0.33
	50-75	134	-0.09	-0.59	0.41	0.78	0.65
	75-100	77	-0.39	-0.54	0.16	0.71	0.67

Low — High

Conclusions

1. When effect sizes and the proportion of nulls are both small, the actual FDR can be a lot smaller than its nominal value (*q*).
2. Even when the FDR is close to its nominal value, the FDP distribution can have extreme skewness opening up the possibility of realizing occasional large FDPs in a real experiments
3. For fixed *q*, depending on ε and m_0/m, skewness and other characteristics of the FDP distribution can be strongly associated with the degree of dependence between test statistics.
4. In explaining the effect of correlated dependent variables on functions of test statistics, one cannot assume that the correlation structure of the test statistics always mimics that of the correlated dependent variables.
5. Consider controlling the *k*-FWER (probability of *k* or more rejections when H_0 is true) in the presence of moderate to extreme dependence, especially when one suspects a large proportion of non-null cases, but with relatively small effect sizes.
6. Please enjoy our online version located at: http://66.43.220.232/james/JSMposter.html

Limitations & Future Directions

1. Our simulations have inherent limitations because it is not possible to investigate all plausible covariance structures.
2. In cases where H_0 was not true, means were set to either a constant or zero.
3. We focused on one multiple testing method (BKY), and primarily one value of the nominal test level (0.10).

Relating Correlation of Variables to Correlation of t-Statistics

Scenario

$\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \mathbf{V})$

n observations of Y and *m t*-tests of

$H_0: \mu_j = 0 \; (j = 1, \ldots m)$

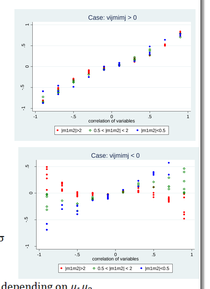
obtain an *m* x 1 vector T of *t*-statistics $T_1, T_2, .., T_m$

Question: How do *n*, **V**, and $\boldsymbol{\mu}$ affect var(T)?
Answer: *n* doesn't matter. $\boldsymbol{\mu}$ and **V** *do* matter.

Illustration: Simulation results: $n = 20$; $m = 2$;
2000 realizations of (T_1, T_2) for each value of μ and ρ.

$$V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \rho = -0.9(0.2)0.9$$
$$\mu = (\mu_1, \mu_{12})' \; \mu_j \sim U(-3,3) \text{ (20 times for each value of } \sigma$$

When $> 0, cor(T_1, T_2) \cong cor(Y_1, Y_2)$
When $< 0, (T_1, T_2)$ can be quite different from $cor(Y_1, Y_2)$ depending on $\mu_1 \mu_2$



Ramifications:
The effect of dependence on the FDP distribution, whether induced by innate correlation between variables or by multiple comparisons in a repeated measures design, is manifested by the correlation structure of the test statistics. In general, it is not true that this correlation structure is the same as that of the original variables. In particular, the correlation structure of the t-statistics that we studied here depends on the interaction between the intra-class correlation over the repeated measures, the covariance structure of the original variables, and the means of the variables. Therefore one cannot assume that simulations with directly generated correlated test statistics provide reliable information as to the effect of correlation between the original variables.

References

1. Gavrilov, Y., Benjamini, Y., and Sarkar, S., (2009). An adaptive step-down procedure with proven FDR control under independence. Annals of Statistics 37: 619 – 629.
2. Benjamini, Y., Krieger, A. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrica* 93: 491–507.
3. Benjamini, Y. and D. Yekutieli. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29: 1165 - 1188.
4. Pawitan, Y., Calza, S., Ploner, A. (2006). Estimation of false discovery proportion under general dependence. *Bioinformatics* 22(24): 3025-3031.
5. Ghosal, S., and Roy, A. Predicting false discovery proportion under dependence. *Under Revision for Publication in J. Amer. Statist. Assoc.*
6. Guo, W. and Sarkar, S. (2010). Stepdown procedures controlling a generalized false discovery rate. *A special volume celebrating the Platinum Jubilee of Indian Statistical Institute*, World Scientific, in press.

Joint Statistical Meetings of the American Statistical Association, July30-August 4, 2011